3-1992

A Compositional Data Approach to the Prediction of Dry Milling Yields

Aziz Bouzaher
Iowa State University

Alicia L. Carriquiry
Iowa State University

A Compositional Data Approach to the Prediction of Dry Milling Yields

Abstract
The yield of products in the dry milling industry is largely determined by the physical properties of the corn kernel. The main objective of this paper is to investigate several statistical models of dry milling yield prediction based on physical characteristics of corn. Data consisting of one hundred corn samples representing a range of genetic traits and quality differences are used. For each corn sample, 16 physical and chemical properties plus six dry milling product yields were measured in a controlled laboratory environment.

For each corn sample, we consider a vector of dry milling product yields and a vector of physical corn characteristics. Several single product models are investigated, two of which implicitly take into account the simplex sample space of product yields. A multivariate model is considered that consists of mapping the sample space from a simplex to unrestricted Euclidean space. Comparison are performed using a jack-knife-like approach.

Keywords
Dry milling, quality characteristics, yield prediction, production function, linear models, compositional data, Cobb-Douglas, translog, continuation ratios, jackknife, multivariate analysis

Disciplines
Agricultural and Resource Economics | Agricultural Economics | Economics

A Compositional Data Approach to the Prediction of Dry Milling Yields

Aziz Bouzaher
and Alicia L. Carriquiry

Working Paper 92-WP 90
March 1992

Center for Agricultural and Rural Development
Iowa State University
Ames, Iowa 50011

Aziz Bouzaher is visiting associate professor of economics, CARD; and Alicia L. Carriquiry is assistant professor of statistics, Iowa State University.

This research has been supported in part by a grant from the Iowa Corn Promotion Board. The authors acknowledge the contributions of Lowell Hill and Marvin Paulsen of the University of Illinois and Allen Kirleis of Purdue University, who all shared in the development of the data set.

Abstract

The yield of products in the dry milling industry is largely determined by the physical properties of the corn kernel. The main objective of this paper is to investigate several statistical models of dry milling yield prediction based on physical characteristics of corn. Data consisting of one hundred corn samples representing a range of genetic traits and quality differences are used. For each corn sample, 16 physical and chemical properties plus six dry milling product yields were measured in a controlled laboratory environment.

For each corn sample, we consider a vector of dry milling product yields and a vector of physical corn characteristics. Several single product models are investigated, two of which implicitly take into account the simplex sample space of product yields. A multivariate model is considered that consists of mapping the sample space from a simplex to unrestricted Euclidean space. Comparisons are performed using a jack-knife-like approach.

Key Words: Dry milling, Quality characteristics, Yield prediction, Production function, Linear models, Compositional data, Cobb-Douglas, Translog, Continuation ratios, Jackknife, Multivariate analysis.

1. INTRODUCTION

The dry milling industry in the United States consumes approximately 160 million bushels of corn annually (USDA 1982; Corn Annual 1991). It is an important link in the food chain between farmers and consumers. The yield of dry milling products is largely determined by the physical properties of the corn kernel. Kernels with a high proportion of hard vitreous endosperm and minimum internal stress cracks provide the highest yield of the more valuable flaking grits. Larger kernels, ease of separation of the germ and endosperm, and a minimum of bran also increase the yield of larger grits. Most of these traits, with the exception of stress cracks, are genetically determined. Although some dry milling firms contract with growers to control variety and handling practices, most continue to buy No.2 corn in the market and to search for measurement technology to determine desirable physical properties. If some easily measurable quality traits are found to be reliable predictors of dry milling yields, corn with those quality traits could be bred and the market would be used to segregate corn on the basis of its potential yield of products.

This paper develops and compares several models for predicting the yield of dry milling products from easily measured physical characteristics. Dry millers can use these measurements to select the corn best suited to meet their contract requirements. The quality of corn to produce maximum grit size differs from corn that produces maximum white goods with fewer flaking grits. Such a model will permit

an economic evaluation of individual quality characteristics that will indicate the premiums that could be paid to farmers for producing different corn varieties. Farmers will in turn encourage plant breeders to invest more research toward corn suited for dry milling.

For the dry milling industry, nonuniform streams of incoming corn, in addition to fluctuations in its intrinsic properties, requires continuous adjustments in mill technology and implies wide variations in the yield of milling products. Identifying the characteristics that determine the yield of primary products could reduce maintenance and set up costs for the dry milling industry and introduce price efficiency in the industrial corn market. Early work by Ladd and Martin (1975) pointed to the importance of not assuming product homogeneity. They developed an economic model for evaluating the current corn-grading system. Manoharkumar et al. (1978) were among the first to seek to relate milling performance and physical and chemical characteristics using laboratory experiments; they reported mainly correlations among the various measurements. Other research identified a positive relationship between density and dry milling yield, and a negative relationship between breakage susceptibility and the yield of dry milling products (Paulsen and Hill 1984; Pomeranz et al. 1986; and Stroshine et al. 1986). However, all this research was essentially confined to revealing important correlations between some corn products and individual physical traits, with no attempt at developing a statistical yield prediction model. Initial investigations of such an approach were conducted by Bouzaher (1987).

The research proposed in this study provides an extension of previous research by simultaneously including all dry milling products and a significant number of measures of quality, using a data set built specifically for this purpose.

The paper is organized as follows. In the second section we present the data set and describe the response variables and the set of potential yield explanatory variables. In the third section we present various univariate models, including two models that attempt to implicitly account for the sample space restriction. The fourth section presents a multivariate approach based on compositional data theory. In the concluding section we discuss the merits of the various models and summarize our findings.

2. DATA DESCRIPTION

A unique data set was collected over a period of two years for the purposes of estimating a model of product yield prediction from measured quality characteristics. In all, 100 samples were collected. Thirty-two samples of flint and dent inbred crosses planted at two locations with a high and low nitrogen application rate were selected to provide a wide range of genetic differences in percent of hard endosperm. An additional ten samples were obtained from superior varieties selected by a dry milling plant. Thirty-nine more samples were provided by a commercial corn breeder, selected to represent a range of genetic characteristics and quality differences related to dry milling. Finally, 19 samples were collected from farmers and elevators, most of

them consisting of a mixture of different varieties and a range of harvesting and drying practices.

A set of 17 physical and chemical tests (Table 1) were performed on each of the samples, in the Agricultural Engineering Laboratory at the University of Illinois: Test Weight (TW,lb/bushel), Wisconsin Breakage (WBT, susceptibility of corn to breakage, %), Stein Breakage (STEIN, a different test for breakage susceptibility, %), Moisture content (MOIST, %) Stress Cracks (SCI, measures extent of high temperature drying on a scale from 1 to 5), Density (DENS, ethanol column test, g/cm^3), Floaters/ sinkers (FLO/SINK, indirect measure of density, %), four Stenvert measurements (based on a grinding resistance test; STIME, time to grind; SCMF, ratio of coarse to medium + fine; SCF, ratio of coarse to fine; S3550, column height at 3550 rpm), Pycnometer (PYCN, another density test, g/cm^3), Starch, Oil, Protein, and Moisture contents by Near Infrared Reflectance (NSTAR, NOIL, NPROT, NMOIST; %), and percent flint (FLINT, percent inbred with dent varieties).

In addition, each sample was dry milled in a short flow pilot mill, at Purdue University's Department of Food Science, in order to obtain a product distribution similar to that obtained from commercial mills. Products were separated by flaking grits, brewers' grits, meal, flour, oil, and feed (Table 2). The yields from each of the six products are reported as the percentage of the total milled corn sample retained on a sieve of a specific mesh size.

Table 3 summarizes all the correlations between product variables and explanatory physical variables [correlation values higher than ABS(.5) are shown in bold]. Similar correlations between the physical

variables reveal, as expected, a high degree of association between several variables, and in particular, the various density measures. Figure 1 presents a three-dimensional scatter plot of the yield data where the variables plotted are percent grits (FG + BG + MEAL), percent flour (FLOUR) and percent oil (OIL). We notice the absence of observations with low flour-low grits and high flour-high grits; this is because of the complementarity between the two types of products within the corn kernel. More details and descriptive analysis of the data set can be found in Hill et al. (1990).

Consider the corn multiproduct yield data in this study. If the yield of each of the six products shown in Table 1 is expressed as the percentage of total yield in each of the 100 corn samples, then the data are a composition, in the sense of Aitchison (1986). A composition consists of observations on the same experimental unit, which are positive and add up to one. Other instances in which data are compositions are household expenditure data, geochemical composition of rocks, and feed rations.

Compositional data have certain characteristics that must be addressed in the statistical analysis. The most important refers to the compositional sample space. Clearly, the appropriate sample space for the elements of a composition is a restricted part of real space called a *simplex*. We define the *simplex* as a set in which each element of the composition is positive, and the sum of all elements equals one. A formal definition of a *simplex* is given in the next section.

Exploratory analyses of these data showed that correlations are high among several quality characteristics. It is therefore expected

that severe multicollinearity problems may arise when the correlated variables are used as predictors in a model, causing an increase in the sampling variance of estimators. The problem of multicollinearity can be addressed in various ways; in this study, we chose a subset of the explanatory variables that did not exhibit high pairwise correlations, recognizing that this is not an in-depth treatment of the problem.

3. UNIVARIATE MODELS

In this section we present five individual product models and discuss their relative predictability.

3.1 Model specification

We let $\mathbf{x}_p$ represent the n x 1 vector whose elements are the yields of dry milling product p (p = 1, ..., D) and $\mathbf{w}_q$ the n x 1 vector with elements equal to the value of quality characteristic q (q = 1, ..., Q). Here, n = 100 observations, D = 6 products and Q = 16 quality traits. The i^{th} observation, then, consists of the vector pair $(\mathbf{x}, \mathbf{w})$. We then consider the following models:

1. Univariate linear model on each x_p (p = 1, ..., D):

$$E(x_p) = \alpha_0 + \alpha_1 w_1 + \ldots + \alpha_Q w_Q \tag{1}$$

2. Restricted Cobb-Douglas on each x_p (p = 1,...,D):

$$E(x_p) = \alpha_0 \cdot w_1^{\alpha_1} \cdot \ldots \cdot w_Q^{\alpha_Q}, \quad \sum_{q=1}^{Q} \alpha_q = 1, \quad \alpha_q > 0 \tag{2}$$

3. Translog model on each x_p (p=1,...,D):

$$E[\log(x_p)] = \alpha_0 + \alpha_1\log(w_1) + \ldots + \alpha_Q\log(w_Q) + \frac{1}{2}\alpha_{11}[\log(w_1)]^2 + \ldots + \frac{1}{2}\alpha_{QQ}[\log(w_Q)]^2 \quad (3)$$

4. Univariate linear model on each log of continuation ratios (following Fienberg 1977):

$$c_1 = x_1,\ c_2 = \frac{x_2}{1 - x_1},\ \ldots,\ c_{D-1} = \frac{x_{D-1}}{1 - \sum_{j=1}^{D-2} x_j} \quad (4)$$

$$E[\log(c_p)] = \alpha_0 + \alpha_1 w_1 + \ldots + \alpha_Q w_Q$$

5. Univariate linear model on logratios y_p ($p = 1,\ldots,D - 1$):

With the logratio transformation: $y_p = \log(\frac{x_p}{x_D})$, $p=1,\ldots,D-1$:

$$E(y_p) = \alpha_0 + \alpha_1 w_1 + \ldots + \alpha_Q w_Q \quad (5)$$

The linear model is the easiest to estimate and interpret. The continuation ratios model and the linear model on logratios were chosen because they implicitly account for the interdependence between product yields, and the restriction in their sample spaces. The Cobb-Douglas and translog models were chosen from a restricted class of functions to test the hypothesis that the relationship between product yields and quality traits can be described in terms of an economic "production function." Physical traits are used as inputs (like labor, capital, and raw materials) that are transformed into dry milling products (see for example, Chalfant 1984; Chambers 1989; Mittelhammer et al. 1981). The interest in describing the underlying production technology by a statistical yield prediction model, if successful, can produce very rich information for further analysis of the existence of a market for

quality traits in corn. The major restrictions in the production function models, which are positive monotonicity and quasi-concavity in the input variables, essentially stipulate that (i) additional units of any input can never decrease the level of output and (ii) as the utilization of a particular input rises, holding all other inputs fixed, the associated marginal increment in output cannot increase.

All models were estimated using SAS Stepwise or SAS GLM. The usual residual diagnostics were performed to verify model validity. Multicollinearity among regressors was tested using a method by Belsley et al. (1980) and by inspection of variance inflation factors. In polynomial models, multicollinearity was reduced by centering regressors, around their mean, and by including a subset of the explanatory variables in each model.

3.2 Model predictability

Predictability of each model was assessed by a jackknife-like approach (Efron 1981). For each model and each product, a predicted value for the i^{th} observation was obtained by fitting the model to the remaining n - 1 observations. The "best" model for each product was the model with the smallest δ, where

$$\delta = \sum_{j=1}^{n} (obs._j - pred._j)^2.$$

The analysis consisted in first estimating 30 separate models (5 model types and 6 products). Very quickly it became clear that, because of the nonindependence between products, no good models were to be obtained for all products separately, and in particular, for brewer's grits and oil. A grits variable was defined (as Grits = FG + BG + MEAL)

to correspond with the total amount of the premium products that are extracted from the vitreous (hard) part of the corn kernel. A summary of the predictability of the best models is given in Table 4 for grits. Similar information was obtained for flour (Table 5).

In both cases, relative model rankings were the same with the "best" model being the translog, closely followed by the general linear model. These two models indicate that the most important physical characteristics common to the prediction of both grits and flour yields are: Stein breakage susceptibility (STEIN), stress cracks index (SCI), pycnometer (PYCN), and NIR-oil (NOIL). Traits that appear to be significant in the prediction of flour alone are: SCF and SCMF (both Stenvert hardness measures). Only one trait appears to be significant in the prediction of grits alone: test weight (TW); this corroborates previous findings (Bouzaher 1987). Surprisingly, none of the four Stenvert tests, designed to measure various aspects of hardness, appears to be significant in the prediction of grits; these tests were shown to be good predictors of hardness by cereal chemists (Pomeranz et al. 1986; Kirleis 1987).

4. A COMPOSITIONAL DATA APPROACH

We now present a different approach to predicting dry milling yield from quality traits. We develop a model based on Aitchison's (1982, 1986) compositional data approach that was used primarily to analyze data pertaining to the geochemical composition of rocks, but is also applicable to any compositional data with "the intrinsic feature that the proportions of the composition are naturally subject to a unit-

sum constraint" (Aitchison 1986, xiii). We first present some relevant theoretical background, largely drawn from Aitchison (1986), before applying the approach to our data set.

4.1 Theory

A D-part composition is defined as a 1 x D vector x, with:

$$x_p > 0,\ p = 1,\ \ldots,\ D,\ \ and\ \sum_{p=1}^{D} x_p = 1\ .$$

In our application, x_p represents the proportion of dry milling products in a given sample. Subcompositions can be defined for any subset of a D-part composition that are then normalized to form new compositions in lower dimensional space. As an example of a subcomposition, consider the one defined as grits. Then a new composition is formed with grits, flour, oil, and feed.

In the preceding section, it was argued that the appropriate space for D-part composition is a simplex, and an informal definition of a simplex was given. The six-part dry-milling product composition is completely determined given knowledge of any five of its products. Here, we give a more precise definition. The sample space for D-part compositions is a d = D - 1 dimensional simplex embedded in a D-dimensional real space. It is the set:

$$S^d = \left[(x_1,\ldots,x_D):\ x_p > 0,\ p = 1,\ \ldots,\ D;\ \sum_{p=1}^{D} x_p = 1\right].$$

Difficulties associated with compositions

An obvious difficulty that is encountered when trying to fit the usual univariate regression models to each of the products is that each

product yield (expressed as a percentage of total yield) must be between zero and one. This clearly makes the usual assumption of normality untenable. Furthermore, univariate modeling of individual product yield may lead to hardly believable predictions, as would be the case if the sum of individually predicted yield percentages were larger than one. These two problems follow from the sample space restriction.

Other difficulties associated with compositional data can be mentioned:

1. The high dimensionality of compositions makes conclusions about the multivariate pattern of variability hard to ascertain. In particular, examination of the data in lower dimensions, by projection, may constitute, at best, a partial analysis. In addition, graphical interpretation of data patterns, as traditionally done in unrestricted Euclidian space, may be highly distorted due to the unit-sum constraint. The multiproduct yield data highlighted in this study, while consisting of only 100 observations, do not easily lend themselves to traditional methods of exploration. Difficulties arise due to the number of elements in the compositions (in this case, six).

2. The absence of an interpretable covariance structure when using the usual covariance or correlation estimates among components of the composition. Three main problems are noted:

(i) Negative bias of correlations.

Since $\sum x_p = 1$, *and since* $Cov(x_p, \sum x_p) = 0$

$$\sum_{j \neq p}^{D} Cov(x_p, x_j) = -Var(x_p) .$$

Thus there will be at least one negative covariance element in each row of the matrix $C = \{Cov(x_p,x_j);\ p,j = 1, \ldots, D\}$, posing serious interpretation problems.

(ii) Subcomposition inconsistencies due to the relationship between the usual covariance matrix of a subcomposition and that of the full composition. The magnitude, sign, and rank ordering of the covariance associated with two specific parts can change erratically as we move from full composition to lower dimensional subcompositions (see Tables 6 and 7).

(iii) Basis difficulty. No relationship between the usual covariance of a composition and the covariance matrix of its basis (e.g., the basis of the dry milling composition is made up of the original product data, in pounds, before it is expressed as a set of proportions).

3. Difficulty of parametric modeling for studying compositional variability patterns, in the absence of "rich" families of distributions over the simplex sample space S^d. Clearly, random variables which are restricted to the interval (0,1) as the elements of a composition are, cannot be assumed to follow a distribution such as the normal distribution. Only the Dirichlet class of distributions, based on independent, equally scaled gamma-distributed components, are defined over S^d. However, Aitchison (1986) points out major limitations of the Dirichlet class for compositional data analysis because "every Dirichlet composition has a very strong implied independence structure" (60).

Because of these difficulties, the following transformation of the original compositional data will enable us to arrive at a more

meaningful analysis of the patterns of variability in any composition in general, and in the dry milling data in particular. The transformation described here maps the data from the simplex into unrestricted Euclidean space, allowing the use of distributions such as the normal distribution as a model. In addition, the richness and flexibility of the multivariate normal family of distributions will be available for linear modeling and hypothesis testing about the relationship between dry milling yield and physical characteristics.

Consider the *Logratio transformation*:

$$y_p = \log(x_p/x_j),\ p \neq j,$$

$$\text{if } x_p,\ x_j \in S^d \quad \text{then} \quad y_p = \log(x_p/x_j) \in \mathbb{R}^d.$$

Covariance structure

The covariance structure of a D-part composition is given by

$$\sigma_{ij.k\ell} = \text{Cov}\{\log(x_i/x_k),\ \log(x_j/x_\ell)\},\ i,j,k,\ell = 1,\ \ldots,\ D,$$

where only $2^{-1}(D\text{-}1)D$ of these covariances can be independently assigned (which is the same number of covariances as in the case of an unrestricted (D-d)-dimensional random vector). These logratio covariances are completely determined by the $2^{-1}(D\text{-}d)$ logratio variances: $\tau_{ij} = \text{var}\{\log(x_i/x_j)\}, i = 1,\ \ldots,\ D-1;\ j = i+1,\ \ldots,\ D,$ where τ_{ij} measures the variability of component x_i relative to component x_j.

In addition, and letting $\xi_{ij} = E\{\log(x_i/x_j)\}$, for a D-part composition, it is possible to construct the *compositional variation*

array. The compositional variation array is defined as the matrix $T = \{\xi_{ij} \backslash \tau_{ij}\}$ with zeros on the diagonal, variances above the diagonal, and means below the diagonal.

Logratio covariance matrix

While the compositional variation array is very useful for describing patterns of compositional variability, it is necessary to be able to fully describe the covariance structure of a composition. Let:

$$\sigma_{ij} = \sigma_{ij.DD} = \mathrm{Cov}\{\log(x_i/x_D), \log(x_j/x_D)\} \text{ for } i,j = 1, \ldots, D-1.$$

The matrix $\Sigma = \{\sigma_{ij}; \text{ for all } i \text{ and } j\}$ is a $(D-1) \times (D-1)$ *logratio covariance matrix*, which determines the covariance structure through the relationships:

$$\sigma_{ij.k\ell} = \sigma_{ij} + \sigma_{k\ell} - \sigma_{i\ell} - \sigma_{jk}.$$

Σ is then the variance-covariance matrix of the $(D-1) \times 1$ vector $y = \{y_i = \log(x_i/x_D)\}$, $i = 1, \ldots, D-1$.

In addition, from the definition of the logratio covariance matrix, we have:

(i) $y \in \mathbf{R}^d$, since the transformation $x \in S^d \rightarrow y = \log(x_{-D}/x_D) \in \mathbf{R}^d$, is one-to-one (where x_{-D} is the vector x without component D).

(ii) The negative bias difficulty is eliminated.

(iii) The basis difficulty is eliminated by the existence of a direct and exact relationship between the covariance structure of any composition and that of its underlying basis.

(iv) Σ is invariant under the group of permutations of the parts of the compositions, thus making any statistical analysis invariant to the choice of the composition anchor or component divisor.

(v) The covariance structure of subcompositions is readily available only in the case of the variation matrix T (use $T_s = STS'$, with S being a selection matrix of 0s and 1s). For the case of the logratio covariance matrix, construction of Σ_s from Σ is possible but nontrivial (see Aitchison 1986, equation 5.24, p. 101).

The additive logistic transformation

Transformations, such as power transformations, are often used to obtain data that are normally distributed. This is, of course, due to the fact that there exists a large battery of procedures that can be easily applied to normally distributed data. We use a transformation presented by Aitchison (1986) termed the *logratio transformation*. The logratio transformation used to resolve the difficulties associated with the usual covariance structure of compositions is also used to find a rich and flexible parametric class of distributions on S^d to study variability patterns in the simplex sample space.

Following Aitchison (1986, p. 113), a D-part composition $\mathbf{x}$ is said to have an *additive logistic normal* distribution $\mathcal{L}^d(\mu, \Sigma)$ when $y = \log(x_{-D}/x_D)$ has an $N^d(\mu,\Sigma)$ distribution (we note that Σ is precisely the logratio covariance matrix defined in the previous section).

We then have available, through this transformation between compositions and logratios, the whole battery of statistical procedures based on multivariate normality, assuming that the logistic normality assumption of compositions is a valid one. In this paper we are of course concerned with linear modeling of the mean to analyze the dependence of product yield composition on physical trait variables.

4.2 Compositional Variability Analysis

Tables 6 and 7 give the usual covariance matrices for the full six-part product yield composition and for the four-part subcomposition obtained by defining a grits component as GRITS = FG + BG + Meal. Inspection of these covariance matrices illustrates the negative bias and the subcomposition inconsistencies discussed previously.

In addition, from the variation array of the six-product composition, given in Table 8, we observe the following:

(i) The largest relative variation between product yields is between FG and Flour with $\tau_{FG,Flour} = .30$; in addition, $\xi_{FG,Flour} = 0.96$ with $\xi_{FG,Flour} > \tau_{FG,Flour}$ indicates that the percentage of FG yield is consistently larger than that of Flour yield (this observation is corroborated by the fact that a large number of the corn samples collected were known to have high density with the potential for high FG yield).

(ii) The smallest relative variation between product yields is between Meal and Flour with $\tau_{Meal,Flour} = 0.019$; in addition, $\xi_{Meal,Flour} = -0.337$ and $\xi_{Meal,Flour} < \tau_{Meal,Flour}$ indicates that not only does Meal yield tend to be smaller than Flour yield, but that is the case for a large number of corn samples. Again, these conclusions are corroborated by inspection of the data.

4.3 Logratio Linear Modeling

We now use the compositional data framework to explore several dry milling yield prediction logratio linear models. To conduct this analysis, we have used the microcomputer software package CODA developed

by Aitchison as a companion to his book on compositional data. However, we should note that an important limitation of this package as currently configured is that it can only handle data sets with a maximum of 10 part-compositions, 10 covariates (explanatory variables), and 100 observations. This limitation can be avoided by using other statistical packages such as SAS, since the analyses on the logratio transformed data is the usual regression-type analyses. Unfortunately, clear, informative graphical analyses, included in CODA are not yet available elsewhere.

Estimation of μ and Σ

With the assumption that the pattern of dry milling yield variability is of $\mathcal{L}^d$ (μ, Σ) form, the estimation of μ and Σ from the logratio data matrix

$Y = [y_1, \ldots, y_d]$ with $\{y_i = \log(x_i/x_D);\ i = 1, \ldots, d = D - 1\}$ is given by:

$$\hat{\mu} = n^{-1}Y'1,$$

$$\hat{\Sigma} = (n-1)^{-1}(Y - \hat{\mu})(Y - \hat{\mu})',$$

where n is the number of observations and z′ denotes the transpose of vector z. For the full six-product composition, we have:

$$E(y) = [0.290,\ 0.733,\ -1.007,\ -0.670,\ -2.208]$$

$$\Sigma = \begin{bmatrix} 2.229 & 0.702 & -0.007 & -0.237 & 0.060 \\ 0.702 & 0.573 & 0.214 & 0.101 & 0.055 \\ -0.007 & 0.214 & 0.265 & 0.183 & 0.102 \\ -0.237 & 0.101 & 0.183 & 0.294 & 0.005 \\ 0.060 & 0.055 & 0.102 & 0.005 & 0.251 \end{bmatrix}.$$

For the grits four-product composition, we have:

$E(y) = [1.349, -0.670, -2.208]$

$$\Sigma = \begin{bmatrix} 7.462 & -0.265 & 0.640 \\ -0.265 & 2.293 & 0.050 \\ 0.640 & 0.050 & 2.515 \end{bmatrix}.$$

Logratio linear models

Let W represent a matrix of covariates, and assume: $f(x|w) \sim \mathcal{L}^d(W\beta, \Sigma)$, then $[y_1, \ldots, y_d] = Y = W\beta + E$, where the rows of the error matrix E are assumed independent and each row is distributed as $N^d(0, \Sigma)$.

To estimate specific models and test hypotheses about various parameterizations, we need to estimate the parameter matrix β and the error logratio covariance matrix Σ. The estimation can be done either by maximum likelihood under the normality assumption or by multivariate least squares.

Let $x = [x_1 = \text{GRITS}, x_2 = \text{Flour}, x_3 = \text{Oil}, x_4 = \text{Feed}]$. Then:

$y = [y_1 = \log(\text{GRITS/Feed}), y_2 = \log(\text{Flour/Feed}), y_3 = \log(\text{Oil/Feed})]$

Tests for normality of marginal distributions of y lead to accepting the underlying model assumptions.

From the results on univariate models, we restrict the set of covariates (regressors) to the most important physical traits in predicting dry milling yield: TW, STEIN, PYCN, SCI, and SCMF. We then specify the following model for the i^{th} observation:

$$[y_1, y_2, y_3] = \alpha_1 + \alpha_2 TW + \alpha_3 STEIN + \alpha_4 PYCN + \alpha_5 SCI + \alpha_6 SCMF + \alpha_7 TW^2 + \alpha_8 STEIN^2 + \alpha_9 \log(TW) + [e_1, e_2, e_3] \quad (6)$$

where α_i (i=1, ..., 9) are (1 x 3) dimensional parameter vectors.

We adopt Aitchison's approach of testing a lattice of hypotheses from this standard model to determine a "best" model; each member of the lattice corresponds to a simple reparametrization of the standard model. The advantage of this approach is that a generalized likelihood ratio test of a hypothesis h within the standard model m is readily available once the residual matrices R_m and R_h are estimated (detailed development of these tests are in Aitchison, 1986, pp. 162-166). Figure 2 (where $|R_h|$ is the residual determinant of model "h" and p_h is the associated significance probability) gives the lattice of hypotheses tested within model (6). Starting at level 1 we reject the hypothesis of random variation with no dependence on quality traits because of a negligible significance probability. At level 2, while the logarithmic hypothesis is also rejected, we cannot reject the linear dependence hypothesis and this gives us the working model:

$$[y_1, y_2, y_3] = \alpha_1 + \alpha_2 TW + \alpha_3 STEIN + \alpha_4 PYCN + \alpha_5 SCI + \alpha_6 SCMF + [e_1, e_2, e_3] \quad (7)$$

with estimated parameter matrix:

$$\alpha_1 = [\ -8.736 \quad -0.104 \quad -1.367\ ]$$

$$\alpha_2 = [\ \ 0.049 \quad \ 0.005 \quad -0.015\ ]$$

$$\alpha_3 = [\ -0.004 \quad \ 0.040 \quad \ 0.006\ ]$$

$$\alpha_4 = [\ \ 5.245 \quad -0.267 \quad -0.235\ ]$$

$$\alpha_5 = [\ \ 0.022 \quad -0.024 \quad \ 0.020\ ]$$

$$\alpha_6 = [\ \ 0.245 \quad -0.425 \quad \ 0.227\ ]$$

and estimated error covariance matrix:

$$e_1 = [\ 2.150 \quad 1.032 \quad 0.442\ ]$$

$$e_2 = [\ 1.032 \quad 1.956 \quad 0.303\ ]$$

$$e_3 = [\ 0.442 \quad 0.303 \quad 2.204\ ].$$

An inspection of the parameter matrix confirms that the yield of premium products GRITS is positively dependent on TW, PYCN, and SCMF, and these physical traits could then be used for yield predictions.

Finally, we note that the linear/logarithmic dependence hypothesis also could not be rejected and could be the basis for another working model.

5. SUMMARY

This paper developed an approach to predict yields of dry milling products from measurable quality characteristics, which could then be used by the dry milling industry to select corn best suited to meet the demand for their products.

We developed several univariate models and discussed their relative merits. We also estimated a lattice of multivariate models based on compositional data analysis, taking explicitly into account the simplex nature of the sample space. We believe this is the first application of this methodology to this type of data.

While a number of other model specifications could be tested within this framework, we hope the emphasis on the methodology would make it useful for the study of other agricultural data sets.

Table 1. Variable Definition of Quality Tests

Variable	Physical Trait/Test Description
TW	Test Weight (lbs./bushel)
WBT	Wisconsin Breakage Test (%)
STEIN	Stein Breakage Test (%)
MOIST	Moisture Content (%)
SCI	Stress Crack Index (1.0 - 5.0 with 5.0 being most severely fractured)
DENS	Density-alcohol Test (grams/cm3
FLO	Floaters Test (%)
STIME	Stenvert time-to-grind (seconds)
SCM	Stenvert coarse/(medium + fine) (ratio)
SCF	Stenvert coarse/fine (ratio)
S3550	Stenvert column height at 3550 r.p.m.
PYCN	Pycnometer test (grams/cm^3)
NSTAR	Starch (%) obtained by NIR
NOIL	Oil (%) obtained by NIR
NPROT	Protein (%) obtained by NIR
NMOIST	Moisture (%) obtained by NIR
FLINT	Degree of inbred flint variety (%)

Table 2. Dry Milled Products Grouped by Wire Size

Product	Variable	Wire/Mesh size
Flaking Grits	FG	3.5 - 5.0
Brewers' Grits	BG	7.0 - 10.0
Meal	MEAL	16.0
Flour	FLOUR	PAN
Oil	OIL	GERM*15%
Hominy Feed	FEED	HULLS+GERM*85%

Table 3. Correlations Between Product Yields and Physical Traits

	FG	BG	MEAL	FLOUR	OIL	FEED
TW	+.58	-.32	-.57	-.48	-.59	-.53
WBT	+.31	+.09	-.13	-.49	-.20	-.53
STEIN	-.38	+.34	+.41	+.29	+.26	+.18
SCI	+.16	+.21	-.04	-.41	-.07	-.37
DENS	+.67	-.30	-.56	-.72	-.43	-.64
FLO	-.72	+.23	+.61	+.83	+.52	+.76
PYCN	+.69	-.23	-.51	-.77	-.52	-.74
STIME	+.86	-.46	-.79	-.88	-.58	-.72
SCMF	+.41	-.02	-.34	-.60	-.23	-.48
SCF	-.40	+.25	+.37	+.36	+.30	+.32
S3550	-.45	+.01	+.36	+.66	+.26	+.54
NSTAR	-.66	+.42	+.72	+.58	+.41	+.49
NOIL	+.81	-.35	-.73	-.85	-.55	-.79
NPROT	+.71	-.55	-.73	-.57	-.52	-.46
NMOIST	-.01	-.12	-.17	+.09	+.13	+.16
FLINT	+.78	-.47	-.62	-.75	-.54	-.67

Table 4. Predictability of Grits Models

Model	F	R^2	Number of regressors	δ	Rank
Linear	66.4	.87	7	.0264	2
Cobb-Douglas	90.7	.82	5	.0370	5
Translog	73.7	.88	9	.0229	1
Cont. Ratios	83.8	.86	6	.0291	3
logratios	56.7	.80	5	.0344	4

Table 5. Predictability of Flour Models

Model	F	R^2	Number of regressors	δ	Rank
Linear	68.8	.85	7	.0064	2
Cobb-Douglas	36.3	.69	6	.0165	5
Translog	68.7	.90	13	.0049	1
Cont. Ratios	21.8	.56	5	.0065	3
logratios	19.9	.64	6	.0078	4

Table 6. Usual Six-Product Covariance Matrix

	FG	BG	Meal	Flour	Oil	Feed
FG	1.000	-0.703	-0.859	-0.860	-0.627	-0.788
BG		1.000	0.503	0.322	0.277	0.199
Meal			1.000	0.792	0.651	0.620
Flour				1.000	0.550	0.800
Oil					1.000	0.566
Feed						1.000

Table 7. Usual Four-Product Covariance Matrix

	GRITS	Flour	Oil	Feed
GRITS	1.000	-0.803	-0.780	-0.958
Flour		1.000	0.396	0.604
Oil			1.000	0.819
Feed				1.000

Table 8. Six-Product Compositional Variation Array

	FG	BG	Meal	Flour	Oil	Feed
FG	0.000	0.140	0.251	0.300	0.236	0.223
BG	-0.443	0.000	0.041	0.067	0.071	0.057
Meal	1.297	1.740	0.000	0.019	0.031	0.027
Flour	0.960	1.403	-0.337	0.000	0.054	0.029
Oil	2.499	2.942	1.201	1.539	0.000	0.025
Feed	0.290	0.733	-1.007	-0.670	-2.208	0.000

Table 9. Four-Product Compositional Variation Array

	GRITS	Flour	Oil	Feed
GRITS	0.000	0.109	0.087	0.075
Flour	2.019	0.000	0.054	0.029
Oil	3.558	1.539	0.000	0.025
Feed	1.349	-0.670	-2.208	0.000

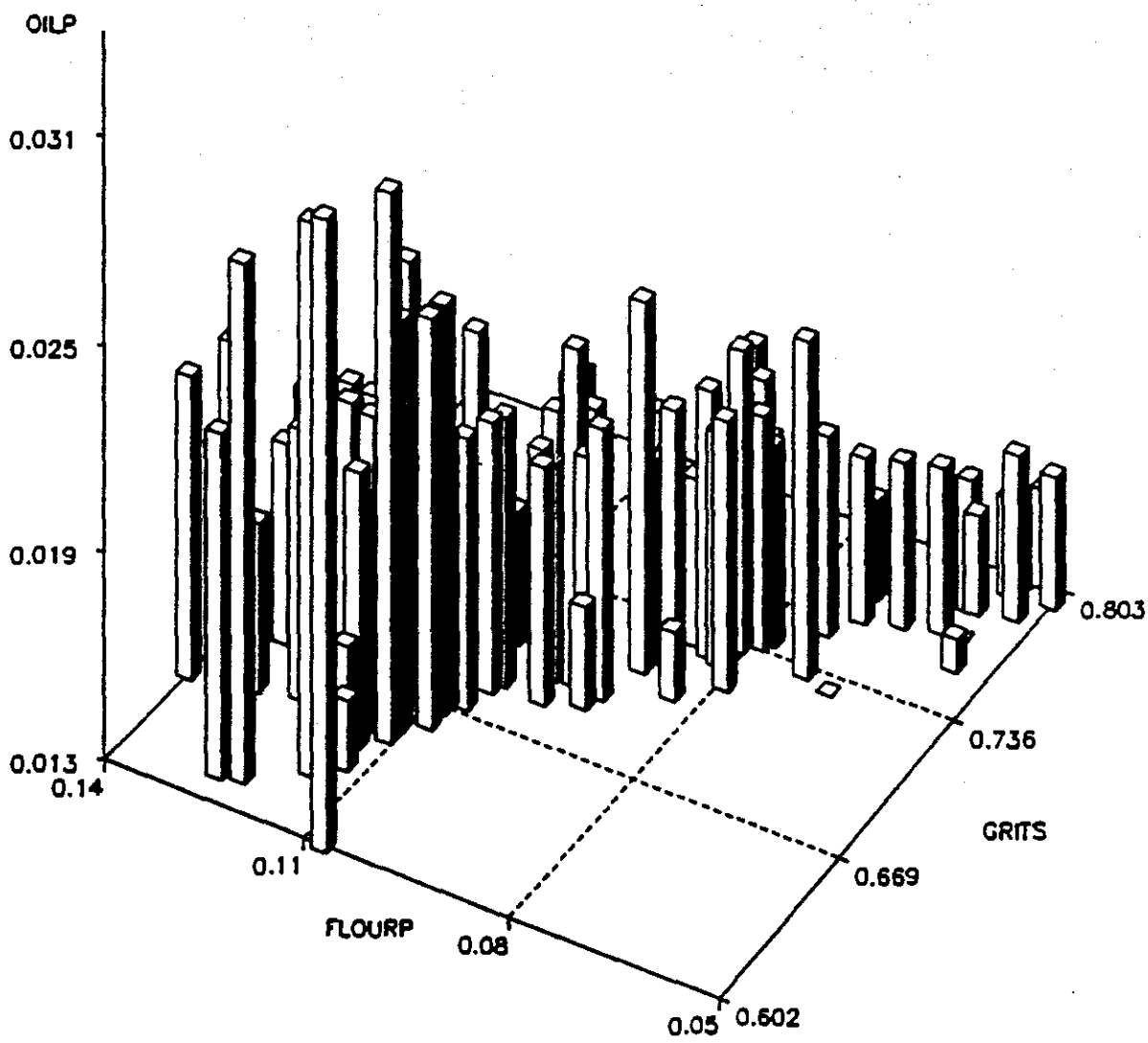

Figure 1. Flour vs Grits vs Oil

Model 1
$|R_m| = 6.277$

Model: Quadratic
$\alpha_2=\alpha_3=\alpha_4=\alpha_5=\alpha_6=\alpha_9=0$
$|R_h| = 29.96$
$p_h < 10^{-5}$

level 3

Model:Linear/Logarithmic
$\alpha_7=\alpha_8=0$
$|R_h| = 6.46$
$p_h = 0.84$

level 2

Model: Linear
$\alpha_7=\alpha_8=\alpha_9=0$
$|R_h| = 6.62$
$p_h = 0.83$

Model: Logarithmic
$\alpha_2=\alpha_3=\alpha_4=\alpha_5=\alpha_6=\alpha_7=\alpha_8=0$
$|R_h| = 32.11$
$p_h = < 10^{-5}$

level 1

Model:
Random variation only
$\alpha_i=0,\ i=2,\ldots,9$
$|R_h| = 6.62$
$p_h < 10^{-5}$

Figure 2. Lattice of Hypotheses for Model (6)

References

Aitchison, J. (1982). "Discussion Paper." J.R. Statist. Soc. B 4: 139-77.

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. New York: Chapman and Hall.

Bouzaher, A. (1987). "Implicit Evaluation of Quality Characteristics in the Dry Milling of Corn." Unpublished Manuscript. Department of Agricultural Economics, University of Illinois at Urbana-Champaign.

Brown, R.S., D.W. Caves, and L.R. Christensen (1979). "Modelling the Structure of Cost and Production for Multiproduct Firms." Southern Economic Journal 16: 256-73.

Chalfant, J.A. (1984). "Comparison of Alternative Functional Forms with Application to Agricultural Input Data." American Journal of Agricultural Economics 66(2):216-20.

Chambers R.G. and R.E. Just. (1989). "Estimating Multioutput Technologies." American Journal of Agricultural Economics 71:980-95.

Corn Annual (1991). Corn Refiners Association, Washington, D.C.

Efron, Bradley (1981). The Jacknife, the Bootstrap and other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics, No. 38.

Fienberg, S.E. (1977). The Analysis of Cross-Classified Categorical Data. Cambridge, Massachusetts: The MIT Press.

Hill, L., M. Paulsen, A. Bouzaher, M. Patterson, A. Kirleis, and K. Bender. (1990). "Economic Evaluation of Quality Characteristics in the Dry Milling of Corn." Preliminary Report, Department of Agricultural Economics, University of Illinois at Urbana-Champaign.

Kirleis, A.W. (1987). Private communication. Department of Food Science, Purdue University.

Ladd, G.W., and M.B. Martin. (1976). "Prices and Demands for Input Characteristics." American Journal of Agricultural Economics 58:21-30.

Manoharkumar, B.P., H. Gerstenkorn, H. Zwingelberg, and H. Bolling. (1978). "On Some Correlations Between Grain Composition and Physical Characteristics to the Dry Milling Performance for Maize." Journal of Food Science and Technology 15(1):1-6.

Mittelhammer, R.C., S.C. Matulich, and D. Bushaw. (1981). "On Implicit Forms of Multiproduct-Multifactor Production Functions." American Journal of Agricultural Economics 63:164-68.

Paulsen, M.R. and L.D. Hill. (1984). "Corn Quality Factors Affecting Dry Milling Performance." Journal of Agricultural Engineering Research 31:255-263.

Pomeranz, Y., Z. Czuchajwski and F.S. Lei (1986). "Comparison of Methods for Determination of Hardness and Breakage Susceptibility of Commercially Dried Corn." Cereal Chemistry 63(1):39-43.

Stroshine, R.L., A.W. Kirleis, J.F. Tuite, L.F. Bauman, and A. Emam (1986). "Differences in Grain Quality Among Selected Corn Hybrids." Cereal Foods World 31(4):311-316.

USDA (1982). U.S. Corn Industry. Agricultural Economic Report No. 479. Economic Research Service, Washington, D.C.